AUTOWORLD: LEARNING MULTI-AGENT TRAFFIC SIMULATION WITH SELF-SUPERVISED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation with realistic traffic agents is essential for validating autonomous driving systems. Existing data-driven simulators learn agent behavior from higher-level abstractions such as 3D bounding boxes and polylines, inferred by upstream perception pipelines. These lossy abstractions discard sensory context that directly shapes agent behavior, limiting the distributional realism that simulation aims to reproduce. To address this limitation, we propose AutoWorld, a traffic simulation framework that grounds agent behavior in raw sensor observations through a self-supervised world model trained on LiDAR occupancy data. Given world model samples, AutoWorld constructs a coarse-to-fine predictive scene context as input to a multi-agent motion generation model. Furthermore, we designed a motion-aware latent supervision objective that enriches AutoWorld's latent representation of scene dynamics. To better exploit this latent space during inference, AutoWorld employs a cascaded Determinantal Point Process framework to guide diversity-aware sampling across both the world model and motion model. Experiments on nuPlan and Waymo demonstrate that AutoWorld improves generation realism, with larger gains in partially observed scenarios where trajectory abstractions are most limited, reducing collision rates by 36%. AutoWorld also remains physically plausible under planner-controlled ego behavior and supports controllable, diverse safety-critical scenario generation.

1 INTRODUCTION

Traffic simulation is essential for developing and testing autonomous driving systems. Data-driven approaches have gained increasing attention for their ability to overcome limitations of heuristic-based simulators. Existing methods Huang et al. (2024); Jiang et al. (2024); Zhou et al. (2024); Wu et al. (2024); Philion et al. (2024); Zhang et al. (2026; 2025); Pei et al. (2026); Christian Rossert (2025) learn agent behavior from motion forecasting datasets Ettinger et al. (2021); Caesar et al. (2020), which represent driving scenes as sequences of 3D bounding boxes and polylines, abstractions inferred by upstream perception pipelines. However, these trajectory abstractions are a lossy compression of the original sensory observations, discarding rich scene information that shapes how real drivers behave, including geometric structure and occlusion context. For example, a driver slowing near a blind intersection responds to raw sensory geometry that no bounding box encodes. Models trained on these abstractions therefore learn the kinematic response without the sensory cause, which limits their ability to reproduce the distributional realism that faithful simulation requires. While Chen et al. (2024) demonstrated that augmenting trajectory data with raw LiDAR observations improves agent behavior generation, how to effectively harness this signal for traffic simulation remains largely unexplored.

Models that forecast future sensor observations, with or without action conditioning, are broadly referred to as "world models" and have proven effective for modeling high-dimensional environments Zhu et al. (2024); Kong et al. (2025). In autonomous driving, video-generative world models have been widely explored, with a primary focus on photorealistic scene synthesis Hu et al. (2023); Gao et al. (2024); Russell et al. (2025); Yang et al. (2024a); Wang et al. (2024b). To favor geometric information, a complementary line of work investigates world modeling through LiDAR data, framed as LiDAR or occupancy forecasting Weng et al. (2021; 2022); Zhang et al. (2023); Liu

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

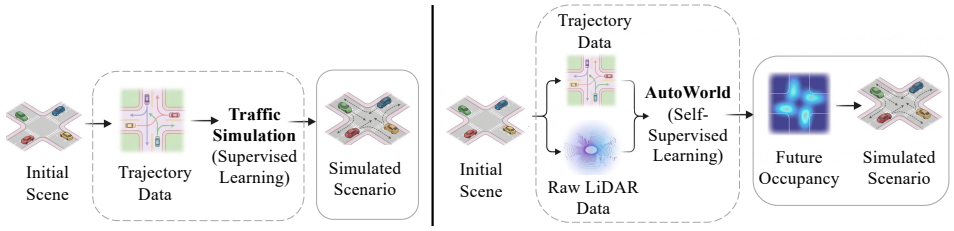

Figure 1: **Comparison of traffic simulation approaches**. (Left) Conventional traffic simulation methods. (Right) AutoWorld grounds simulation in future scene occupancies predicted by a self-supervised LiDAR world model, preserving sensory context discarded by trajectory representations.

et al. (2025). Yet, these methods typically rely on semantic occupancy prediction, which requires costly dense annotation, and apply uniform supervision across spatial regions, underemphasizing the motion dynamics most relevant to behavior generation for traffic simulation.

In this work, rather than learning agent behavior solely from trajectory abstractions, we propose to ground traffic simulation in a world model learned directly from raw LiDAR observations (See Fig. 1). Our key insight is that a self-supervised world model of the driving environment can capture the geometric and dynamic context that trajectory data discards, providing a richer foundation for realistic behavior generation. To this end, we introduce **AutoWorld**, a traffic simulation framework that conditions a multi-agent motion generation model on world model predictions, transferring the world model's sensor-grounded understanding of scene dynamics into the simulation process.

AutoWorld trains a latent LiDAR world model with a motion-aware supervision objective that prioritizes dynamic scene elements without using semantic labels, allowing fully self-supervised training. We then propose a conditional diffusion policy that leverages future latent occupancies and a predictive scene context, encapsulating the anticipated occupancy sequence, to model the joint distribution of future agent behaviors. By doing so, AutoWorld is equipped with a coarse-to-fine context derived from sensory observation. Notably, agent identity is not provided during world model learning; the world model captures only occupancy dynamics, while agent-specific behaviors are produced by the diffusion model, allowing the two components to play complementary roles. Furthermore, to better exploit the latent representations of both models, we propose a cascaded latent sampling strategy applied at inference time without additional training. This strategy promotes structured exploration across two complementary levels: the scene level, guiding sampling of future occupancies from the world model, and the agent level, driving variation in the diffusion-based motion generation. We formalize this using a flow space determinantal point process (DPP) Kulesza & Taskar (2012) with a quality-weighted kernel that balances and sample realism, enabling AutoWorld to cover the space of plausible scene evolutions and agent behaviors without modifying training.

In summary, our **main contributions** are: (1) We propose AutoWorld, a traffic simulation framework that grounds traffic simulation in raw sensor observations through a self-supervised LiDAR world model, addressing the fundamental limitations of trajectory-based approaches. (2) We introduce a motion-aware latent supervision objective that emphasizes dynamic scene elements, enabling self-supervised world model training without semantic labels. (3) We propose a cascaded latent sampling strategy based on a quality-weighted DPP that promotes structured exploration of the joint latent space of both models at inference. (4) Extensive experiments on nuPlan and Waymo demonstrate AutoWorld's competitive generation realism, with larger gains under partial observability. AutoWorld remains reactive to planner-induced ego behaviors and enables controllable and diverse safety-critical scenario generation, while ablations validate its key components.

2 Related Work

We describe two related areas: (1) Traffic Simulation and (2) World Models in Autonomous Driving.

Traffic Simulation. Traffic simulation models the distribution of multi-agent behaviors in dynamic driving environments. Recently, data-driven approaches have gained significant attention as they can overcome limitations of heuristic-based simulators such as poor adaptability to ego-vehicle behavior. Prior work spans CVAEs Igl et al. (2023); Suo et al. (2021); Xu et al. (2022), Transformer-based architectures Girgis et al. (2021); Ngiam et al. (2021); Shi et al. (2024); Zhou et al. (2023), diffusion models Huang et al. (2024); Jiang et al. (2023); Lu et al. (2024); Pronovost et al. (2023); Zhong et al.

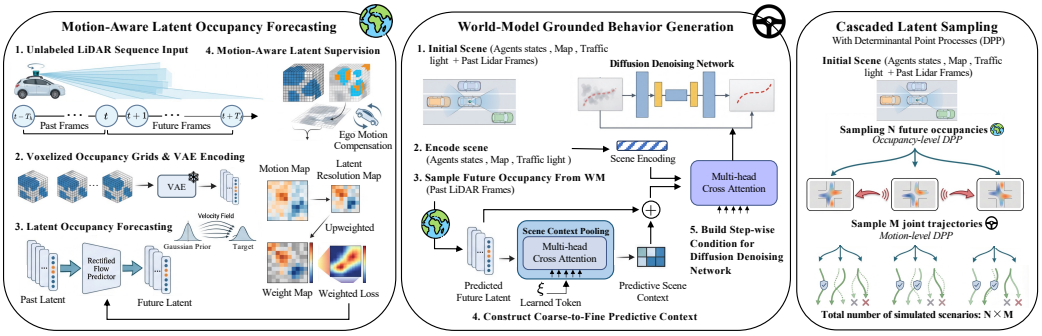

Figure 2: **Overview of AutoWorld**. AutoWorld learns latent scene dynamics from unlabeled LiDAR occupancy sequences and uses predicted future occupancies to condition diffusion-based motion generation. At inference, a cascaded sampling strategy further improves use of the learned latent spaces across both scene evolution and agent behavior without additional training.

(2023a;b), and more recently autoregressive next-token formulations for joint trajectory generation Wu et al. (2024); Zhang et al. (2026; 2025); Zhou et al. (2024); Philion et al. (2024); Peng et al. (2026). Despite architectural differences, these methods learn behavior solely from trajectory abstractions, which discard sensory cues shaping real driving behavior Chen et al. (2024). In contrast, AutoWorld grounds behavior generation in raw sensor observations through self-supervised world models, mitigating the information bottleneck of trajectory-only approaches.

World Models in Autonomous Driving. World models represent and forecast the evolution of driving environments. One line of work uses future sensor prediction for end-to-end driving, either jointly optimized with planning Pan et al. (2022); Hu et al. (2022); Li et al. (2024); Zheng et al. (2024) or used for pretraining perception encoders Min et al. (2023); Yang et al. (2024b); Min et al. (2024). Another line develops generative neural simulators that produce future video Hu et al. (2023); Wang et al. (2024a;b), LiDAR Weng et al. (2021; 2022); Zhang et al. (2023); Liu et al. (2025), or multimodal observations Bogdoll et al. (2025) for synthesis, counterfactuals, or imagination-based planning. These methods mainly emphasize fidelity or controllability, often decoupled from behavior modeling. Rather than treating the world model as a standalone generator, AutoWorld integrates its predictions into motion generation, providing sensor-grounded scene context for traffic simulation.

3 AUTOWORLD

AutoWorld grounds traffic simulation in raw sensor observations by building simulation around a world-modeling formulation. Instead of modeling future agent behavior in isolation, AutoWorld models the joint future distribution of sensor observations and agent states, factorized into a self-supervised LiDAR world model and a diffusion-based motion generation model. This factorization makes world modeling the interface through which raw sensor data informs traffic simulation: the world model learns predictive scene dynamics from high-dimensional unlabeled LiDAR data, while the motion model retains structured trajectory-based inputs for agent-centric behavior generation.

Specifically, we first learn a motion-aware latent occupancy world model in a fully self-supervised manner from unlabeled LiDAR data to forecast future latent occupancies (Sec. 3.2). From these predictions, we build a coarse-to-fine representation of the future scene to condition the diffusion-based motion generation model. (Sec. 3.3). Finally, to promote structured exploration of plausible multi-agent behaviors, we introduce a cascaded latent sampling strategy applied at inference time without additional training (Sec. 3.4). The overall framework is illustrated in Fig. 2.

3.1 PROBLEM FORMULATION.

Given past states of A agents over T_h and contextual information, traffic simulation predicts their future states over T_f. At timestep t, agent states are $s_t = [s_t^1, \ldots, s_t^A]$, where $s_t^i = (x_t^i, y_t^i, v_t^i, \theta_t^i)$ denotes the 2D position, speed, and yaw of agent i. We denote the history by $S_{t-T_h:t} = [s_{t-T_h}, \ldots, s_t]$, with LiDAR observations $O_{t-T_h:t}$ available over the same horizon. Context includes an HD map $V \in \mathbb{R}^{v \times d_v}$ with up to v vector points of dimension d_v, and traffic-light states

$R \in \mathbb{R}^{r \times d_r}$ for up to r lights with feature dimension d_r. Actions are $a_t = [a_t^1, \ldots, a_t^A]$, where $a_t^i = (\dot{v}_t^i, \dot{\theta}_t^i)$ denotes acceleration and yaw rate. Our flow-matching and diffusion timesteps, denoted k_μ and k_f, are distinct from the state-sequence timestep t.

3.2 Motion-Aware Latent Occupancy Forecasting

Latent Occupancy Forecasting. We formulate our world model as a LiDAR-based occupancy predictor operating in the latent space of voxelized occupancies encoded by a VAE. We assume access to additional unlabeled LiDAR sequences $O_{t-T_h:t+T_f}$, which are used to train the latent occupancy predictor. Let $Z_{t-T_h:t}$ denote past latent occupancies and $Z_{t+1:t+T_f}$ the future sequence to be predicted. We build upon Liu et al. (2025), which adopts a rectified flow formulation Liu et al. (2022). We denote the conditioning information as $c^\mu = \{Z_{t-T_h:t}, \tau^{\text{ego}}_{t-T_h:t}\}$, where $\tau^{\text{ego}}_{t-T_h:t}$ denotes the ego vehicle's past trajectory embedding. The model learns a time-dependent velocity field $\mu_\theta(z_t, t, c^\mu)$ that transports samples from a Gaussian prior toward a target future latent occupancy $\tilde{z}$. During training, we sample flow matching timestep $k_\mu \sim \mathcal{U}(0,1)$ and $z_0 \sim \mathcal{N}(0, I)$, construct the interpolation $z_t = (1-t)z_0 + t\tilde{z}$, and optimize the rectified flow objective in Eq. 1:

$$\mathcal{L}^{\text{world}}(\theta) = \mathbb{E}_{k_\mu,\, z_0,\, \tilde{z}} \left\| \mu_\theta(z_{k_\mu}, k_\mu, c^\mu) - (\tilde{z} - z_0) \right\|_2^2 . \tag{1}$$

Motion-Aware Latent Supervision. Occupancy grids are spatially imbalanced. Static structures dominate the volume while dynamic objects occupy sparse but behaviorally critical regions. Uniform flow supervision therefore biases learning toward static structure and underweights motion-driven evolution — an issue amplified in our fully unsupervised setting, which lacks the semantic labels that prior methods use to implicitly separate dynamic from static content.

To address this, we introduce a motion-aware supervision strategy derived directly from temporal occupancy changes. Let Y_t and $Y_{t+\Delta}$ denote non-semantic occupancy grids with validity masks M_t and $M_{t+\Delta}$. After compensating for ego motion by warping Y_t into the frame of $t+\Delta$, producing $\tilde{Y}_{t\to t+\Delta}$, we compute voxel-wise occupancy transitions:

$$C_{t,\Delta}(v) = \mathbf{1}[Y_{t+\Delta}(v) \neq \tilde{Y}_{t\to t+\Delta}(v)] \cdot \tilde{M}_{t\to t+\Delta}(v) \cdot M_{t+\Delta}(v),$$

where $\tilde{M}_{t\to t+\Delta}$ is the validity mask M_t warped to the frame of $t+\Delta$. We downsample this motion map to the latent resolution to obtain $\bar{C}_{t,\Delta} \in \mathbb{R}^{H\times W}$ and construct a latent weight map $W_t(i,j) = 1 + \lambda \bar{C}_{t,\Delta}(i,j)$, where $\lambda > 0$ controls the strength of motion emphasis. The resulting motion-aware rectified flow objective is

$$\mathcal{L}^{\text{world}}(\theta) = \mathbb{E}_{k_\mu, z_0, \tilde{z}} \sum_{i,j} W_t(i,j) \left\| \mu_\theta(z_{k_\mu}, k_\mu, c^\mu)_{i,j,:} - (\tilde{z} - z_0)_{i,j,:} \right\|_2^2 ,$$

where W_t is normalized to unit spatial mean. This formulation preserves the standard rectified flow objective while biasing gradient magnitude toward temporally changing regions.

3.3 World-Model-Grounded Traffic Simulation

We formulate traffic simulation as conditional diffusion over joint future agent trajectories in the action space. Let $\tau_0 = a^{1:A}_{t+1:t+T_f}$ denote the clean future action trajectory over the prediction horizon T_f. The diffusion model generates trajectories by initializing from Gaussian noise $\tau_{K_f} \sim \mathcal{N}(0, I)$ and progressively reversing a predefined noising process over K_f steps. This reverse process defines a conditional Markov chain $p_\psi(\tau_{k_f-1} \mid \tau_{k_f}, c^f)$, where c^f denotes the conditioning information. At each denoising step $k_f = K_f, \ldots, 1$, the transition is:

$$p_\psi\left(\tau_{k_f-1} \mid \tau_{k_f}, c^f\right) := \mathcal{N}\left(\tau_{k_f-1}; f_\psi\left(\tau_{k_f}, k_f, c^f\right), \Sigma_{k_f}\right),$$

where f_ψ and Σ_{k_f} are the reverse-transition mean predictor and variance schedule, respectively.

World-model Conditioning. Given the world-model rollout $\hat{Z}_{t+1:t+T_f} = [\hat{z}_{t+1}, \ldots, \hat{z}_{t+T_f}]$, we construct a compact predictive scene context by temporally subsampling the predicted latent occupancies and aggregating them through multi-head cross-attention pooling ("MHCA") with a learned token, defined as $g = \text{MHCA}(\text{Q} = \xi, \text{K} = \text{V} = \phi_1(\{\hat{z}_{t+e\delta}\}_{e=1}^{\lfloor T_f/\delta \rfloor}))$, where $\phi_1(\cdot)$ indicates a

learnable linear projection. The resulting context g summarizes the forecasted occupancies, where each latent occupancy encodes a predicted scene configuration without explicit agent identities.

Next, we condition the diffusion denoiser on both local and global predictive cues, forming a coarse-to-fine guidance mechanism for multi-agent behavior generation. Specifically, for each future timestep $t+r$, where $r = 1, \ldots, T_f$, we combine the corresponding per-step latent occupancy $\hat{z}_{t+r}$ with the global predictive scene context g. As such, the timestep-specific diffusion condition is defined as $c^f_{t+r} = \mathrm{MHCA}\left(\mathrm{Q} = h,\ \mathrm{K} = \mathrm{V} = \phi_2\left([\hat{z}_{t+r} \oplus g]\right)\right)$, where h denotes the scene encoding obtained from a query-centric Transformer Shi et al. (2024). The $\oplus$ indicates concatenation along the token dimension, and $\phi_2(\cdot)$ maps the concatenated latent occupancies into a shared space. The full diffusion condition is therefore $c^f = \{c^f_{t+r}\}_{r=1}^{T_f}$, which combines historical scene context with predicted future occupancies. This allows the diffusion model to generate agent-specific behaviors grounded in past observations while being guided by the world-model.

3.4 Cascaded Latent Sampling

To better exploit the learned latent spaces, we introduce a cascaded latent sampling strategy that applies Determinantal Point Processes (DPPs) Kulesza & Taskar (2012); Morshed & Boddeti (2025) at both the world-modeling and behavior-generation stages. At inference, this mechanism promotes structured exploration by discouraging similar latent samples, yielding diverse yet plausible future scene evolutions and agent behaviors.

Determinantal Point Processes. Let $\mathcal{X} = \{\hat{x}^{(i)}\}_{i=1}^{K_s}$ be the ground set of generated candidates, and let $\hat{x}^i \in \mathbb{R}^d$. Let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric positive semidefinite kernel (e.g. cosine similarity), and define the kernel matrix $\Lambda \in \mathbb{R}^{K_s \times K_s}$ by $\Lambda_{ij} = \kappa(\hat{x}^{(i)}, \hat{x}^{(j)})$. Following the standard L-ensemble DPP Kulesza & Taskar (2012) formulation on a finite ground set, we define a DPP-based probability mass for the K_s jointly generated samples, i.e., treating $\mathcal{X}$ as the ground set, we evaluate the DPP probability of selecting the full subset as: $\mathcal{P}_\kappa(\mathcal{X}) = \frac{\det(\Lambda)}{\det(\Lambda+I)}$. $\mathcal{P}_\kappa$ increases when samples are diverse and approximately linearly independent, and decreases as redundancy increases, becoming 0 when $\mathcal{X}$ contains any exact duplicates under κ (equivalently, when Λ is rank-deficient). This provides a principled continuous measure of diversity. However, maximizing diversity alone may push samples away from the data manifold. To balance diversity and plausibility, we incorporate quality-aware modulation Morshed & Boddeti (2025) by assigning each sample a quality weight $q^{(i)} \in (0, 1]$ and defining $\kappa_q(\hat{x}^{(i)}, \hat{x}^{(j)}) = q^{(i)}\kappa(\hat{x}^{(i)}, \hat{x}^{(j)})q^{(j)}$. This yields the quality-balanced kernel $\Lambda^q = \mathrm{diag}(q)\Lambda\,\mathrm{diag}(q)$ and the quality-aware DPP probability $\mathcal{P}^q_\kappa(\mathcal{X}) = \frac{\det(\Lambda^q)}{\det(\Lambda^q+I)}$, whose maximization balances diversity and sample quality.

DPP-Guided Sampling. To jointly generate K_s diverse candidates $\left\{\hat{x}^{(i)}\right\}_{i=1}^{K_s}$, we evaluate $\mathcal{P}^q_\kappa$ over the predicted target outputs and incorporate a repulsive guidance term proportional to $\nabla \log \mathcal{P}^q_\kappa$ into the flow ODE during sampling. This term encourages candidate trajectories to spread across distinct modes of the predictive distribution while maintaining high-quality solutions. The guidance is applied only at inference time and remains independent of the underlying training objective.

We apply diversity-guided sampling at both world modeling and behavior generation through a two-stage cascaded inference procedure. First, we sample N future latent occupancies from the world model $\mathcal{X}^{\text{world}} \doteq \{\hat{Z}^{(i)}_{t+1:t+T_f}\}_{i=1}^{N}$, where each latent rollout represents a distinct plausible future occupancy development. We quantify rollout quality via the Mahalanobis distance of the one-step latent estimate to a Gaussian fit of the training data. Let $\mathcal{P}^{q,\text{world}} \doteq \mathcal{P}^q_\kappa(\mathcal{X}^{\text{world}})$. The diversity-guided world-model sampling dynamics follow the continuous-time flow formulation given by

$$\hat{z}^{(i)}_{k_\mu - 1} = \mu_\theta(\hat{z}^{(i)}_{k_\mu}, k_\mu, c^\mu) - \gamma_\mu(k_\mu)\nabla_{z^{(i)}_{k_\mu}} \log \mathcal{P}^{q,\text{world}}_{k_\mu}, \quad i = 1, \ldots, N,$$

where $\gamma(.)$ controls the strength of diversification. Next, conditioned on the step-aware representation derived from each sampled latent representation, we generate M agent behavior trajectories using our conditional diffusion model, $\{\tau^{(i,j)}\}_{j=1}^{M}$. Since behavior generation proceeds via discrete reverse diffusion steps, we apply diversity guidance as a gradient shift to the reverse-step mean at each timestep k_f. Let $\mathcal{P}^{q,\text{motion}} \doteq \mathcal{P}_{\kappa_q}(\mathcal{X}^{\text{motion}})$. The diversity-guided sampling dynamics for the

Table 1: **Simulation realism and physical plausibility on nuPlan.** Realism measures the distributional deviation from logged driving statistics (longitudinal/lateral acceleration magnitude and jerk); other metrics assess physical plausibility. D-dir. Viol. denotes driving-direction violation.

Model	Realism(↓)	Collision(↓)	Off-road Time(↓)	D-dir. Viol.(↓)	Kinematic(↑)
CTG Zhong et al. (2023b)	0.584	27.79	25.22	16.59	79.37
GUMP Hu et al. (2024)	0.472	23.81	30.57	20.52	73.48
SceneD. Qian et al. (2024)	0.355	36.93	<u>14.02</u>	<u>12.59</u>	89.35
D. Policy Chi et al. (2025)	0.604	33.68	19.84	21.00	88.29
Nexus Zhou et al. (2025)	<u>0.318</u>	27.19	16.03	15.64	84.41
OMEGA Li et al. (2026b)	0.348	<u>23.91</u>	14.68	19.48	<u>90.63</u>
AutoWorld (Ours)	**0.266**	**8.14**	**5.93**	**7.31**	**94.89**

behavior trajectory simulation, similar to the dynamics for the world model, becomes

$$\tau_{k_f-1}^{(i,j)} = f_\psi(\tau_{k_f}^{(i,j)}, k_f, c^{f(i)}) - \gamma_f(k_f)\,\nabla_{\tau_{k_f}^{(i,j)}} \log \mathcal{P}_{k_f}^{q,\text{motion}} \qquad j = 1, \ldots, M,$$

where $c^{f(i)}$ denotes the conditioning signal derived from the i-th world-model rollout. For notational simplicity, we omit the stochastic noise term from the variance schedule. Trajectory embeddings are obtained from the scene encoder over agent states. At the behavior level, quality weights encode trajectory-space realism constraints, including collisions, kinematic infeasibility, and off-road violations Gulino et al. (2023), preventing diversity from favoring physically invalid rollouts. Overall, cascaded diversity-guided inference produces $N \times M$ diverse and plausible rollouts.

4 Experiments

Our experiments focus on five key questions: **Q1:** Does AutoWorld generate *realistic* traffic behavior that matches real-world human driving? (Sec 4.1) **Q2:** What benefits does sensor-grounded world modeling provide over trajectory-based simulation? (Sec 4.2) **Q3:** Can AutoWorld exhibit realistic *reactiveness* to planner-controlled ego behavior? (Sec 4.3) **Q4:** Beyond realism and reactiveness, can AutoWorld generate *diverse* and *controllable* behaviors? (Sec 4.4) **Q5:** How do AutoWorld's key design choices contribute to its overall performance? (Sec 4.5)

Experimental Design. AutoWorld is trained on **nuPlan** Karnchanachari et al. (2024) and **Waymo** Sun et al. (2020). We follow Occ3D-Waymo Tian et al. (2023) and NuPlan-Occ Li et al. (2026a) to construct voxel-wise occupancy grids for world-model training. AutoWorld predicts 8 seconds ahead from 2 seconds of history on nuPlan and 1 second on Waymo in closed loop, following the Open Sim Agents benchmark. We use receding-horizon closed-loop simulation at 1 Hz for both occupancy and motion. The world model is rolled out autoregressively without access to future ground-truth occupancy. Notably, Waymo provides LiDAR data *only for the observation horizon* in the test split. For DPP sampling, we use a cosine-similarity kernel at both levels with a power-cosine schedule Gao et al. (2023), yielding weaker guidance early and stronger guidance toward the end of sampling. We follow each benchmark's realism criteria. On nuPlan, we measure the Wasserstein distance between simulated and logged driving-profile distributions, together with log-independent physical-plausibility metrics, including collision rate, off-road time (fraction of simulation time spent off-road), driving-direction violation rate, and kinematic feasibility rate. On Waymo, we follow the Open Sim Agents protocol and report likelihood-based realism metrics covering kinematic behavior, agent interactions, and map adherence, their aggregate Realism Meta Metric (RMM), and minADE. Thus, realism metrics are benchmark-specific and not directly comparable across datasets.

4.1 Comparison to State of the Art

We begin by evaluating AutoWorld on nuPlan for distributional realism against logged human driving and physical plausibility, comparing against recent and reproduced traffic-generation methods in Table 1. AutoWorld achieves state-of-the-art performance across all reported metrics, improving both realism and physical plausibility over prior diffusion-based methods. Notably, it reduces collision by nearly 16 percentage points over the next-best method (8.14 vs. 23.91), while reducing off-road time (5.93 vs. 14.02 for the next best), showing that the collision reduction does not come at the expense of road adherence. We further evaluate AutoWorld on the Waymo Open Sim Agents benchmark, a complementary evaluation focused on matching the distribution of human

Table 2: **Simulation realism under the Waymo Open Sim Agents evaluation protocol.** RMM (Realism Meta Metric) is a weighted aggregation of kinematic, interactive, and map-based metrics and serves as the primary ranking metric.

Training	Method	RMM(↑)	Kinematic(↑)	Interactive(↑)	Map-based(↑)	minADE(↓)
RL Post-training	RLFTSim Ehsan Ahmadi (2025)	0.7857	0.4927	0.8129	0.9183	1.3252
	SMART-R1 Pei et al. (2026)	0.7858	0.4944	0.8110	0.9201	1.2885
	DecompGAIL Guo et al. (2026)	0.7864	0.4919	0.8152	0.9176	1.4209
	LLM2AD Wang et al. (2025)	0.7779	0.4846	0.8048	0.9109	1.2827
Imitation Learning	SMART Wu et al. (2024)	0.7814	0.4854	0.8089	0.9153	1.3931
	CAT-K Zhang et al. (2025)	0.7846	**0.4931**	0.8106	0.9177	1.3065
	TrajTok Zhang et al. (2026)	0.7852	0.4887	0.8116	**0.9207**	1.3179
	SceneStreamer Peng et al. (2026)	0.7731	0.4493	0.8084	0.9127	1.4252
	AutoWorld (Ours)	**0.7865**	**0.4931**	**0.8143**	0.9185	**1.3051**

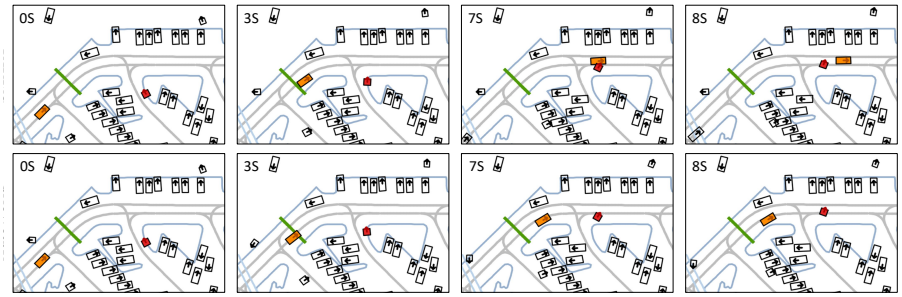

Figure 3: **Qualitative rollout comparison.** Trajectory-baseline (top) leads to a collision between the vehicle (■) and pedestrian (■), while AutoWorld (bottom) maintains a realistic yielding interaction.

driving behavior. Table 2 groups methods by training paradigm, following Guo et al. (2026). AutoWorld is trained solely with imitation objectives and uses no reward-based post-training. Among imitation-trained simulators, AutoWorld achieves the highest realism (RMM 0.7865) and matches RL-fine-tuned simulators without reward-based fine-tuning. See Figure 3 for qualitative examples.

4.2 Benefit of Sensor-Grounded Simulation

To study when raw sensor data provides information beyond lossy trajectory abstractions, we consider occlusion as a representative test case Chen et al. (2024), following the protocol of Lange et al. (2024). Both Waymo and nuPlan provide high-quality agent tracks from offline auto-labeling pipelines Ettinger et al. (2021); Karnchanachari et al. (2024), allowing trajectories to persist even when agents are unobservable by ego sensors. We estimate occlusion using BEV line-of-sight ray casting from the ego vehicle and count, for each valid agent, timesteps where it is geometrically occluded but present in the trajectory labels. We compute a scene-level occlusion score and form equal-sized **Low-**, **Moderate-**, and **High-Occlusion** subsets with comparable agent densities. We compare AutoWorld against Nexus Zhou et al. (2025), a recent diffusion-based SOTA with public code and checkpoints, and two controlled diffusion baselines.Traj.-Diffusion predicts joint multi-agent actions from trajectory context only, while LiDAR-Diffusion additionally incorporates LiDAR observations, fused with HD-map features following DeTra Casas et al. (2024). Figure 4 reveals a clear sensitivity of trajectory-based simulation to partial observability on nuPlan. As occlusion increases, Nexus and Traj.-Diffusion degrade substantially in realism and physical plausibility. Directly incorporating LiDAR alleviates this degradation, reducing collision rate from 58.01% to 42.68% over Traj.-Diffusion under high occlu-

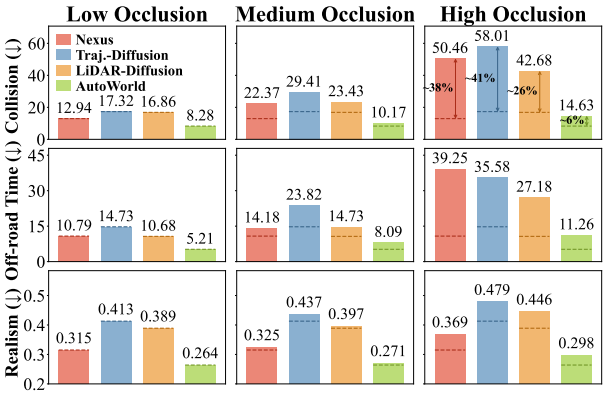

Figure 4: **Simulation performance across occlusion levels** Dashed lines denote low-occlusion performance as reference for degradation at higher occlusion levels.

sion. AutoWorld further reduces collisions to 14.63%—*about 36 percentage points below Nexus*—while achieving a realism score of 0.298 and substantially less degradation across occlusion levels. These results indicate that sensory information is valuable beyond trajectory abstractions, and that *explicitly modeling sensory-world dynamics* is more effective than direct sensor conditioning.

4.3 Closed-Loop Reactiveness

The evaluations in Sec. 4.1 let the behavior model control both the ego and surrounding agents. To test reactiveness beyond this self-controlled setting, we instead use an independent ego planner while the simulator controls surrounding agents. We consider both rule-based and learning-based planners. As shown in Tab. 3, AutoWorld remains robust under externally controlled ego behavior, consistently outperforming Nexus across both planners. Compared with its self-controlled setting, AutoWorld retains high kinematic feasibility (94.89→89.38/88.05) while maintaining substantially lower collision and off-road rates than Nexus. This demonstrates that AutoWorld can respond safely and feasibly when ego behavior is generated independently of the simulator.

Table 3: **Closed-loop reactiveness under planner-controlled ego behavior on nuPlan.** The ego vehicle is controlled by an external planner, while all other agents are generated by the simulator. E2A/A2A denote ego-to-agent and agent-to-agent collision rates, respectively.

Model	Collision(↓)		Off-road Time(↓)	Kinematic(↑)
	E2A	A2A		
Rule-based Planner				
Nexus	34.51	22.02	17.93	81.59
AutoWorld	10.48	8.93	12.69	89.38
Learning-based Planner Chi et al. (2025)				
Nexus	38.29	27.24	21.83	74.46
AutoWorld	13.05	11.32	13.25	88.05

4.4 Diverse and Controllable Simulation

A useful simulator should support both *diverse* and *controllable* generation. We probe these properties through safety-critical scenario generation, explicitly steering the simulator toward challenging interactions with a downstream planner. Diffusion models naturally support inference-time guidance, unlike the less explicit control of next-token prediction methods. We therefore investigate whether AutoWorld can generate targeted safety-critical interactions while preserving its simulation realism. Following Safe-Sim Chang et al. (2024), we evaluate on nuPlan with a planner in the loop and apply the same safety-critical guidance objective to AutoWorld and Nexus. The objective steers a designated adversarial agent toward collision with the planner-controlled ego while discouraging other-agent collisions and off-road behavior. As shown in Table 5, AutoWorld achieves a higher adversarial collision rate than Nexus (41.84 vs. 36.12), while substantially reducing unintended collisions (7.03 vs. 23.68) and off-road time (3.60 vs. 13.92). We further assess diversity through the variance of collision properties across generated outcomes, following Safe-Sim. DPP increases diversity across all three attributes, particularly relative speed (1.27→4.58) and collision point (1.53→6.19), while improving controllability and realism.

Figure 5: **Controllable and diverse safety-critical scenario generation on nuPlan.** Target Collision denotes ego–adversarial-agent collision, while Non-Target Collision denotes all other collisions. Off-road Time is measured for the adversarial agent, and Relative Speed is the ego–adversarial speed at collision. AutoWorld achieves a higher target collision rate while substantially reducing non-target collisions; DPP sampling further increases collision diversity.

Model	Collision Metrics				Collision Diversity (Var)(↑)		
	Target Collision(↑)	Non-Target Collision(↓)	Off-road Time(↓)	Realism(↓)	Angle (rad)	Relative Speed (m/s)	Point (m)
Nexus	36.12	23.68	13.92	0.437	2.41	1.04	1.86
AutoWorld-DPP	34.56	9.42	5.28	0.289	1.93	1.27	1.53
AutoWorld+DPP	**41.84**	**7.03**	3.60	**0.274**	3.77	**4.58**	**6.19**

4.5 Ablation Studies

Motion-Aware Latent Supervision. We study whether motion-aware latent supervision improves dynamic scene forecasting without semantic labels and whether these gains transfer to traffic simulation. We compare semantic supervision (Sem-RF), uniform non-semantic supervision (NSem-RF), and our motion-aware non-semantic supervision (NSem-MA-RF) on Waymo. Table 4 shows that

Table 4: **Effect of motion-aware latent supervision on Waymo.** The MA-RF objective improves world modeling and translates into gains in traffic simulation without requiring semantic labels. Except for FVD(10^{-3}) and minADE, higher values are better.

	World model			Traffic simulation	
Model	**FVD**	**IoU$_{st}$**	**IoU$_{dy}$**	**RMM**	**minADE**
Sem-RF	23	0.89	0.73	0.7651	1.4081
NSem-RF	51	0.85	0.58	0.7547	1.4826
NSem-MA-RF	35	0.87	0.71	0.7618	1.4334

Table 5: **Ablation study on world-model grounding and DPP sampling strategies on Waymo**. The design choices provide complementary gains, with full AutoWorld achieving the best performance. Higher values are better.

Model	**RMM**	**Kinematic**	**Interactive**	**Map-based**
AutoWorld	0.7746	0.4878	0.7968	0.9099
w/o mo-DPP	0.7648	0.4846	0.7926	0.8892
w/o occ-DPP	0.7563	0.4833	0.7758	0.8871
w/o DPP	0.7597	0.4840	0.7839	0.8861
w/o WM	0.7472	0.4749	0.7741	0.8682

uniform supervision performs reasonably on static regions but underperforms on dynamic ones, indicating a bias toward dominant static occupancy. Motion-aware supervision mitigates this bias, raising dynamic IoU to 0.71 while preserving static IoU$_{st}$ (0.87), and improves rollout fidelity (FVD: 51 vs. 35). These world-model gains translate to simulation, improving both realism and trajectory accuracy over uniform supervision (RMM: 0.7547 to 0.7618), nearly matching the semantic baseline in RMM. Overall, motion-aware supervision provides a label-free alternative that strengthens dynamic forecasting and closes much of the gap to semantic supervision.

World-model Grounding and Cascaded Sampling. Table 5 shows that world-model grounding is the primary contributor to AutoWorld's performance. Removing the world model (WM) leads to the largest degradation across all metric groups, reducing RMM from 0.7746 to 0.7472 while also lowering kinematic, interactive, and map-based scores. Diversity sampling further provides consistent gains on top of world-model grounding. Removing motion-level DPP (mo-DPP) reduces interactive realism ($0.7968 \rightarrow 0.7926$), whereas removing occupancy-level DPP (occ-DPP) causes a larger drop in overall realism ($0.7746 \rightarrow 0.7563$). Notably, removing both DPP modules still outperforms the variant without the world model, indicating that predictive scene grounding contributes more substantially than sampling diversity alone. Combining world-model grounding with cascaded DPP sampling yields the best performance. Ablations use data from the Waymo interactive val split.

Scaling with Additional Data. Finally, we investigate whether sensor-grounded world modeling provides a more scalable path for leveraging additional data. We compare against the trajectory-diffusion baseline from Sec. 4.2, augmented to match AutoWorld's parameter count. The baseline converts additional data into trajectory supervision, while AutoWorld uses the corresponding LiDAR observations to incrementally train its world model. As shown in Fig. 6, AutoWorld improves more consistently with data, while the trajectory baseline quickly saturates. This suggests that world modeling better exploits sensor data by transforming raw observations into predictive scene context. While AutoWorld's gains also diminish at larger scales, stronger world models and tighter integration of scene dynamics with behavior generation may enable further scaling, pointing to future directions.

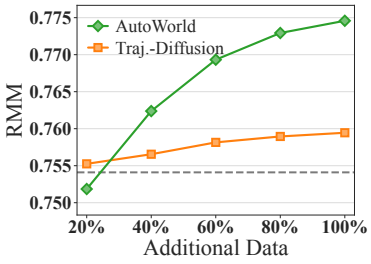

Figure 6: **Scaling with different data modalities on Waymo.** Dashed line denotes Traj.-Diffusion without additional data.

5 Conclusion

We introduced **AutoWorld**, a world-model-grounded framework for traffic simulation that uses raw LiDAR observations to complement structured trajectory representations. Across nuPlan and Waymo, AutoWorld improves simulation realism and physical plausibility, with larger gains under partial observability. Additional experiments showed that AutoWorld reacts consistently to planner-induced ego behaviors and supports controllable safety-critical scenario generation, while diverse sampling produces varied yet realistic agent behaviors. We further demonstrated more favorable scaling with additional raw sensor data than a trajectory-only baseline. Our results highlight sensor-grounded world modeling as a promising direction for realistic and scalable traffic simulation beyond trajectory abstractions. Future work can explore stronger multimodal world models, where cameras provide complementary semantic and appearance cues for more realistic and human-like behavior, as well as tighter coupling between sensory-world prediction and behavior generation.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring disclosure under the ICLR AI Policy. We used generative AI tools only for editing the paper to improve readability and for creating and editing parts of the software code, both of which are tasks with recommended disclosure. All AI-assisted work was reviewed by the authors: AI-edited text was checked and revised for accuracy and clarity, and AI-generated or AI-assisted code was tested and verified for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed implementation and training settings, hyperparameters and default values, dataset preprocessing, and evaluation protocols in the supplementary material. We also describe the architectures and experimental configurations used for all baselines and ablations to facilitate reproduction of our results. We will publicly release the source code upon acceptance of this paper to facilitate reproduction of our results.

ETHICS STATEMENT

This work uses only publicly available datasets and involves no new data collection from human subjects. We have no conflicts of interest or sponsorships that could have influenced this research.

REFERENCES

Daniel Bogdoll, Yitian Yang, Tim Joseph, Melih Yazgan, and J Marius Zollner. Muvo: A multimodal generative world model for autonomous driving with geometric representations. In *2025 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2020.

Sergio Casas, Ben Agro, Jiageng Mao, Thomas Gilles, Alexander Cui, Thomas Li, and Raquel Urtasun. Detra: A unified model for object detection and trajectory forecasting. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Wei-Jer Chang, Francesco Pittaluga, Masayoshi Tomizuka, Wei Zhan, and Manmohan Chandraker. Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In *European conference on computer vision (ECCV)*. Springer, 2024.

Kan Chen, Runzhou Ge, Hang Qiu, Rami Ai-Rfou, Charles Qi, Xuanyu Zhou, Zoey Yang, Scott Ettinger, Pei Sun, Zhaoqi Leng, et al. Womd-lidar: Raw sensor dataset benchmark for motion forecasting. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2025.

Lukas Brostek Christian Rossert, Johannes Drever. combot: an ensemble combination model combining results from smart-tiny-clsft with a cognitive behavior mode. *CVPR Workshop on Autonomous Driving (WAD)*, 2025.

Hunter Schofield Ehsan Ahmadi. Rlftsim: Multi-agent traffic simulation via reinforcement learning fine-tuning. *CVPR Workshop on Autonomous Driving (WAD)*, 2025.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision (CVPR)*, 2021.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. *arXiv preprint arXiv:2104.00563*, 2021.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ke Guo, Haochen Liu, Xiaojun Wu, and Chen Lv. Decompgail: Learning realistic traffic behaviors with decomposed multi-agent generative adversarial imitation learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. *Advances in Neural Information Processing Systems*, 35:20703–20716, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Yihan Hu, Siqi Chai, Zhening Yang, Jingyu Qian, Kun Li, Wenxin Shao, Haichao Zhang, Wei Xu, and Qiang Liu. Solving motion planning tasks with a scalable generative model. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Zhiyu Huang, Zixu Zhang, Ameya Vaidya, Yuxiao Chen, Chen Lv, and Jaime Fernández Fisac. Versatile scene-consistent traffic scenario generation as optimization with diffusion. *arXiv preprint arXiv:2404.02524*, 2024.

Maximilian Igl, Punit Shah, Paul Mougin, Sirish Srinivasan, Tarun Gupta, Brandyn White, Kyriacos Shiarlis, and Shimon Whiteson. Hierarchical imitation learning for stochastic environments. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Max Jiang, Yijing Bai, Andre Cornman, Christopher Davis, Xiukun Huang, Hong Jeon, Sakshum Kulshrestha, John Lambert, Shuangyu Li, Xuanyu Zhou, et al. Scenediffuser: Efficient and controllable driving simulation initialization and rollout. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, et al. Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, et al. 3d and 4d world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 2012.

Bernard Lange, Jiachen Li, and Mykel J Kochenderfer. Scene informer: Anchor-based occlusion inference and trajectory prediction in partially observable environments. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

Bohan Li, Xin Jin, Hu Zhu, Hongsi Liu, Ruikai Li, Jiazhe Guo, Kaiwen Cai, Chao Ma, Yueming Jin, Hao Zhao, et al. Scaling up occupancy-centric driving scene generation: Dataset and method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In *European conference on computer vision*. Springer, 2024.

Shihao Li, Naisheng Ye, Tianyu Li, Kashyap Chitta, Tuo An, Peng Su, Boyang Wang, Haiou Liu, Chen Lv, and Hongyang Li. Optimization-guided diffusion for interactive scene generation. In *European Conference on Computer Vision (ECCV)*. Springer, 2026b.

Tianran Liu, Shengwen Zhao, and Nicholas Rhinehart. Towards foundational lidar world models with efficient latent flow matching, 2025. URL https://arxiv.org/abs/2506.23434.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Jack Lu, Kelvin Wong, Chris Zhang, Simon Suo, and Raquel Urtasun. Scenecontrol: Diffusion for controllable traffic scene generation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Uniworld: Autonomous driving pre-training via world models. *arXiv preprint arXiv:2308.07234*, 2023.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15522–15533, 2024.

Mashrur M Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021.

Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. *Advances in neural information processing systems*, 2022.

Muleilan Pei, Shaoshuai Shi, and Shaojie Shen. Advancing multi-agent traffic simulation via r1-style reinforcement fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Scenestreamer: Continuous scenario generation as next token group prediction. In *International Conference on Learning Representations (ICLR)*, 2026.

Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nick Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. *Advances in Neural Information Processing Systems*, 2023.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI conference on artificial intelligence*, 2024.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020.

Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mingyi Wang, Jingke Wang, Tengju Ye, Junbo Chen, and Kaicheng Yu. Do llm modules generalize? a study on motion generation for autonomous driving. *Conference on Robot Learning (CoRL)*, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*. Springer, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, and Nicholas Rhinehart. Inverting the pose forecasting pipeline with spf2: Sequential pointcloud forecasting for sequential pose forecasting. In *Conference on Robot Learning (CoRL)*, 2021.

Xinshuo Weng, Jiayuan Nan, K-H Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, and Kris M Kitani. S2net: Stochastic sequential pointcloud forecasting. In *European Conference on Computer Vision (ECCV)*, 2022.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Danfei Xu, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Bits: Bi-level imitation for traffic simulation. *arXiv preprint arXiv:2208.12403*, 2022.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*, 2023.

Zhejun Zhang, Peter Karkus, Maximilian Igl, Wenhao Ding, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Closed-loop supervised fine-tuning of tokenized traffic models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Zhiyuan Zhang, Xiaosong Jia, Guanyu Chen, Qifeng Li, and Junchi Yan. Trajtok: What makes for a good trajectory tokenizer in behavior generation? In *International Conference on Learning Representations (ICLR)*, 2026.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*. Springer, 2024.

Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. In *Conference on robot learning (CoRL)*, 2023a.

Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023b.

Yunsong Zhou, Naisheng Ye, William Ljungbergh, Tianyu Li, Jiazhi Yang, Zetong Yang, Hongzi Zhu, Christoffer Petersson, and Hongyang Li. Decoupled diffusion sparks adaptive scene generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Zikang Zhou, HU Haibo, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.